NIST
Translation Evaluation Progress & Plan

GALE Kick-off Meeting

Mark Przybocki, Greg Sanders, Audrey Le, John Garofolo
Alvin Martin, Christophe Laprun

NIST Speech Group

http://www.nist.gov/speech/tests/gale

September 27-28, 2005
San Jose, California

Outline

- Proof-of-Concept Exercises
  - POC-1 review/wrap-up (*Text*, *Arabic MT04)*
  - POC-2 results (*Text*, *Chinese MT05)*
  - POC-3 status (*Audio, Ara+Chi new data)*
- NIST Post Editing Interface
- Translation Dry-Run Evaluation Proposal
- Remaining Issues

Proof-of-Concept #1 - Review/Wrap-up

- Post Editing Arabic system translations from MT04. text data
- Presented at the GALE Eval. and Data meeting in July '05
  - Online: http://nist.gov/speech/tests/gale/poc/doc/gale-poc1-v31.pdf
- Goals:
  - Test post editing concepts
  - Use to develop evaluation protocols
- High level summary of findings from POC-1:
  - Post editor agreement showed promise
  - The editors on average handled about ***780*** words per hour *(or about 2 newswire docs)*
  - System rankings were stable with various methods of counting edits, and correlated with human assessments
  - Estimated that 30 newswire documents may suffice to differentiate +- %5 absolute differences in system performance with 95% confidence
    - Based on the mean and variance across documents, using score averages across 5 editors

Proof-of-Concept #2

- Goals:
  - Repeat the exercise with what is believed to be a more difficult data set (Chinese) due to poorer system translations
  - Address lessons learned from POC-1 (where relevant)
    - More documents
  - Use POC-2 to prepare for the Translation Dry-Run evaluation
    - Does rate of post editing change with different data set?
    - Can we use more editors editing less documents?
    - Any special issues arise from translations of Chinese text?

POC-2 – Data Set from MT05

- Documents
  - Chinese newswire text
  - 25 MT05 documents
    - The set selected for human assessment in the NIST 2005 MT evaluation
  - 272 segments, ~7600 reference words
- Reference
  - NIST adjudicated the four MT05 references into one Gold-Standard
    - Where we found ambiguities across the references we asked two native Chinese speakers to help resolve the differences
- System output that was Post Edited
  - Two top performing MT05 systems
    - GOOGLE – 22%, ISI – 20% (BLEU, on this 25 doc set against GS ref)

Comparing POC Exercises 1 and 2

	POC-1	POC-2
Changed between Exercises		
Source Lang.	Arabic (MT-04)	Chinese (MT-05)
Test set size	10 documents	25 documents#
Systems	3 varied performance (31%, 20%, 17% BLEU)	2 top systems (22% and 20% BLEU)
Post Editors	5 each edited all 30 docs	12 each edited 10 docs##
Unchanged between Exercises		
Source data	Text	
Guidelines	Only slight modifications	
Edit Interface	Only slight modifications	

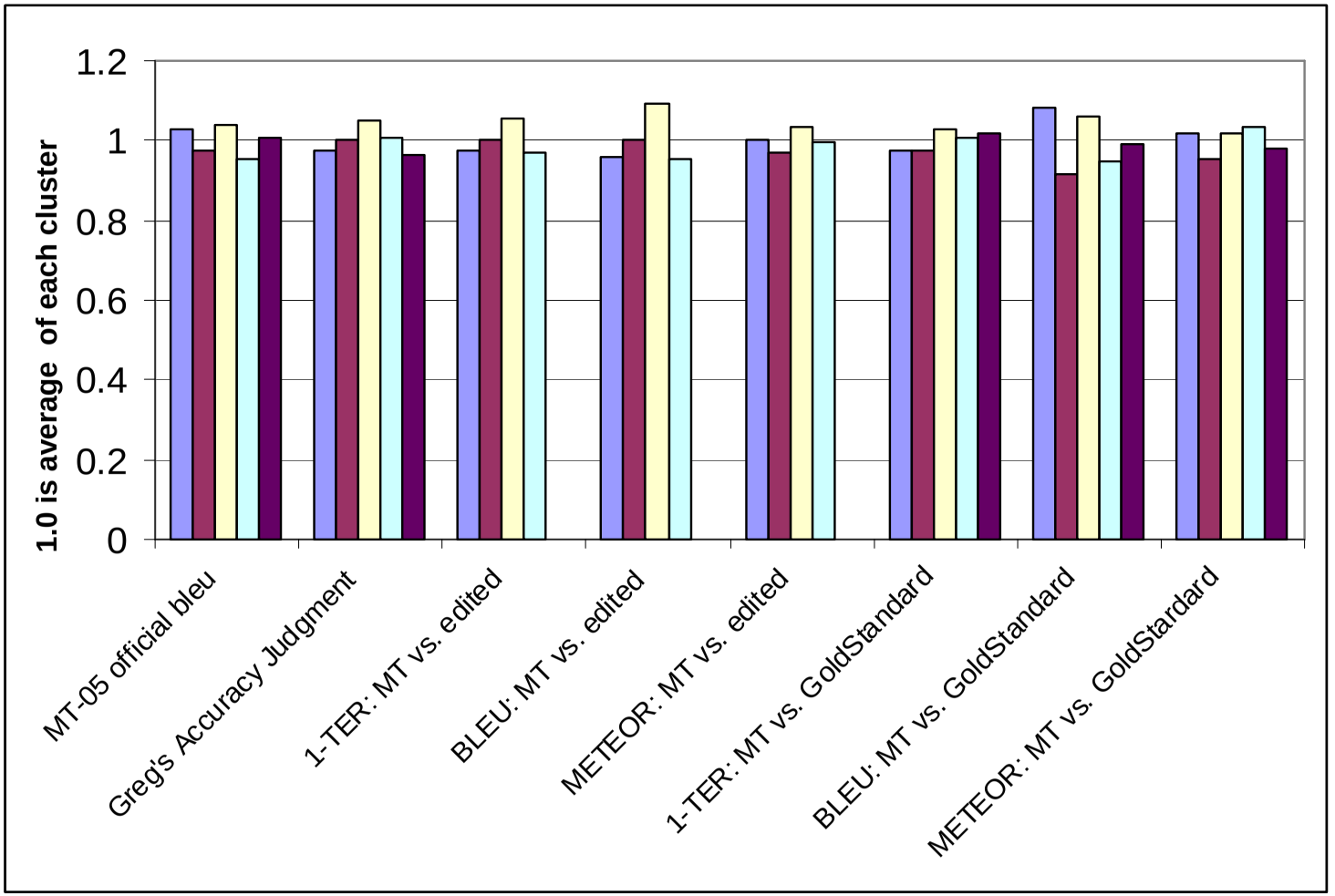
Every segment of each document has human assessment scores
Each document edited by at least two editors

POC-2: Post Editing

- Editors:
  - Mostly NIST volunteers with no previous post-editing experience
  - Provided with guidelines and a few documented examples
- POC-2 Datasets (25 docs for 2 systems = 50 docs to be edited):
  - The 50 document translations were divided into 5 sets of 10, each set contained 5 ISI and 5 GOOGLE document translations
    - Sets were chosen to have approximately equal average BLEU score
    - Each set was given to two editors, order of document presentation was reversed between them
- Editing
  - The post editing paradigm permits an editor to concentrate on a single segment without looking for “meaning” ahead or behind.
    - Text data has an imposed one-to-one segment mapping between reference translation and system translation

POC-2 DataSets Equivalent

(values shown are normalized)



ANOVA results consistent with equivalent difficulty as measured by various metrics shown

POC-2: Rate of Editing

12 Volunteers with no training

Post Editor	Resulting edited segment same as reference	Same color implies same dataset but doc order varies	Approximate	
			Time editing	Words/Hour
L.C.	*21 (108)*		n/a	
V.D.	*18 (108)*		n/a	
Greg Sanders	*1 (103)*		n/a	
B.L.	*5 (103)*		9hr 00min	370
Alvin Martin	*2 (109)*		7hr 40min	435 **(530)**
English Teacher	*2 (114)*		5hr 00min	645 **(785)**
Jon Fiscus	*3 (98)*		4hr 35min	795
M.C.	*0 (98)*		4hr 25min	815
E.M.	*3 (114)*		3hr 55min	820
Wade Shen	*1 (103)*		3hr 45min	890
K.R.	*1 (114)*		3hr 20min	960
Doug Jones	*2 (109)*		2hr 20min	1380

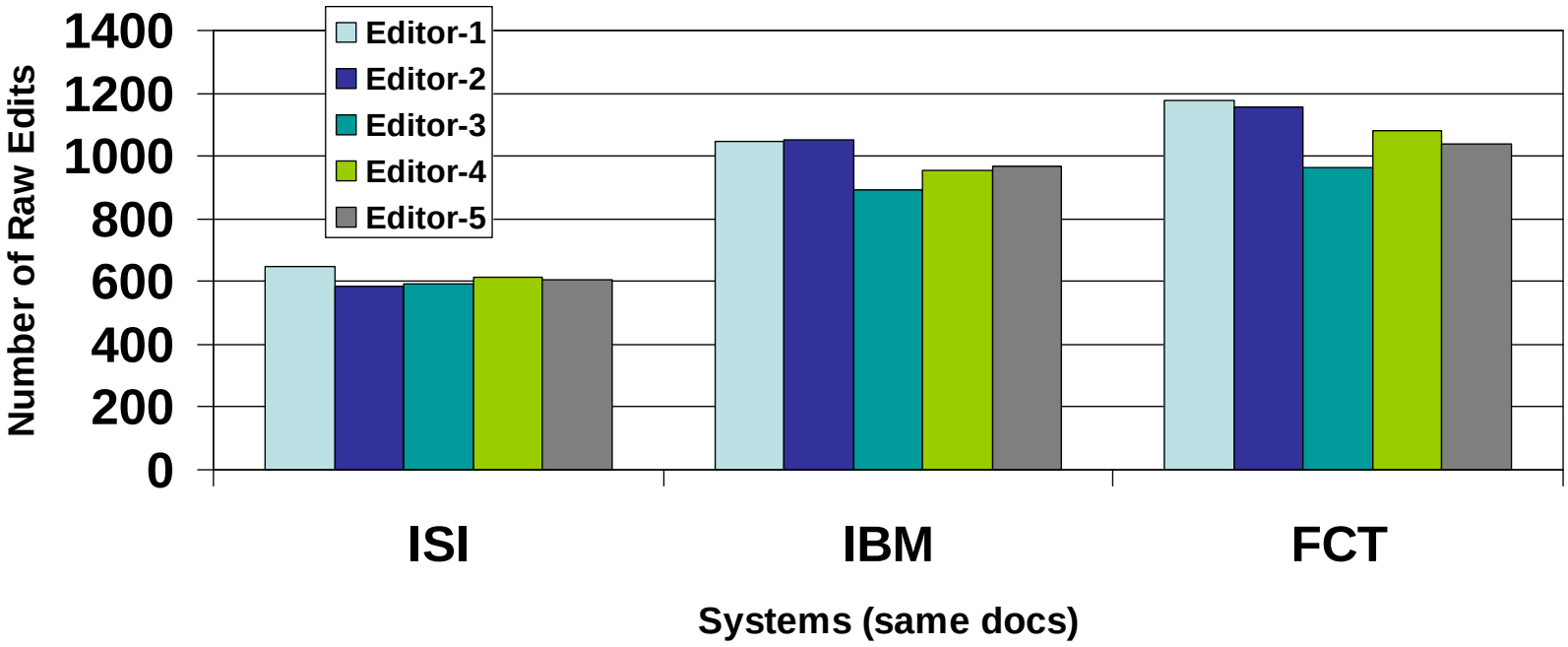
Rate for Arabic in POC1

POC-2: Metrics

- NIST calculated performance using various metrics
  - BLEU (IBM)
    - Weighted n-gram co-occurrence measure for n-grams 1-4
  - Meteor (CMU)
    - Weighted measure of precision and recall of word matches
    - Stemming and synonymy are used to find additional matches
  - WER (NIST/sclite)
    - Word Error Rate, traditional ASR metric
  - **TER** (UMD/BBN)
    - Translation Error Rate, measure of edit distance, is similar to WER but counts block moves as a single error
- Using
  - The final ***edited MT output*** as reference
  - The original ***unedited MT output*** as test (hypothesis)
  - For TER & Meteor the Gold Standard token count was used as the denominator
    - BLEU and WER use the token count from the edited MT

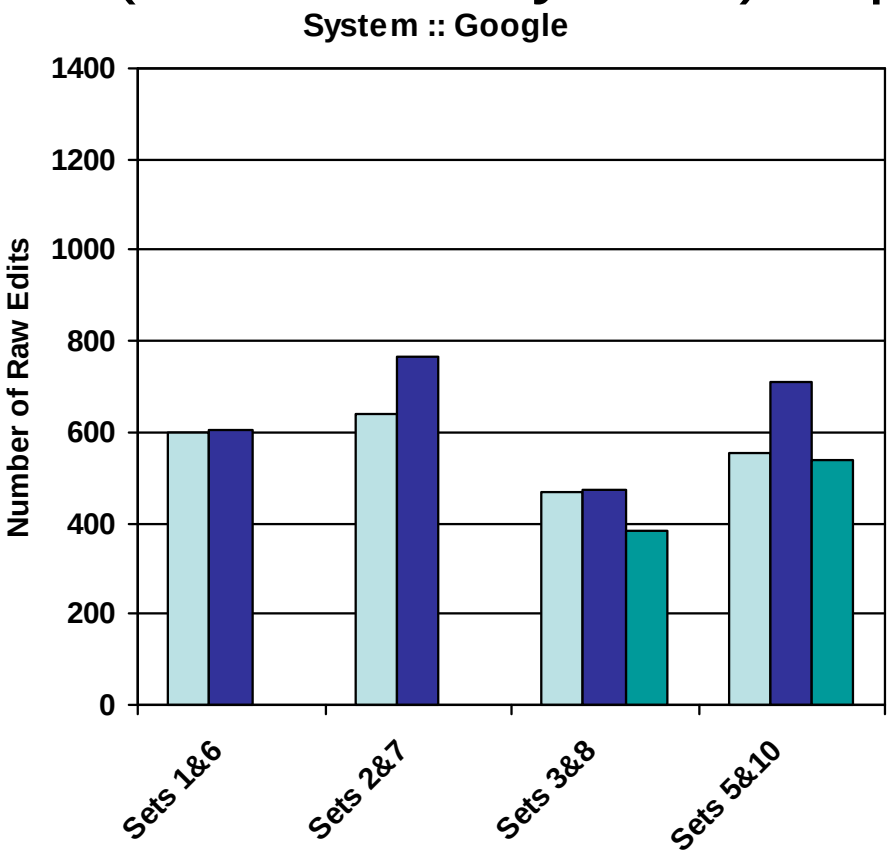
POC-1: Editor Agreement

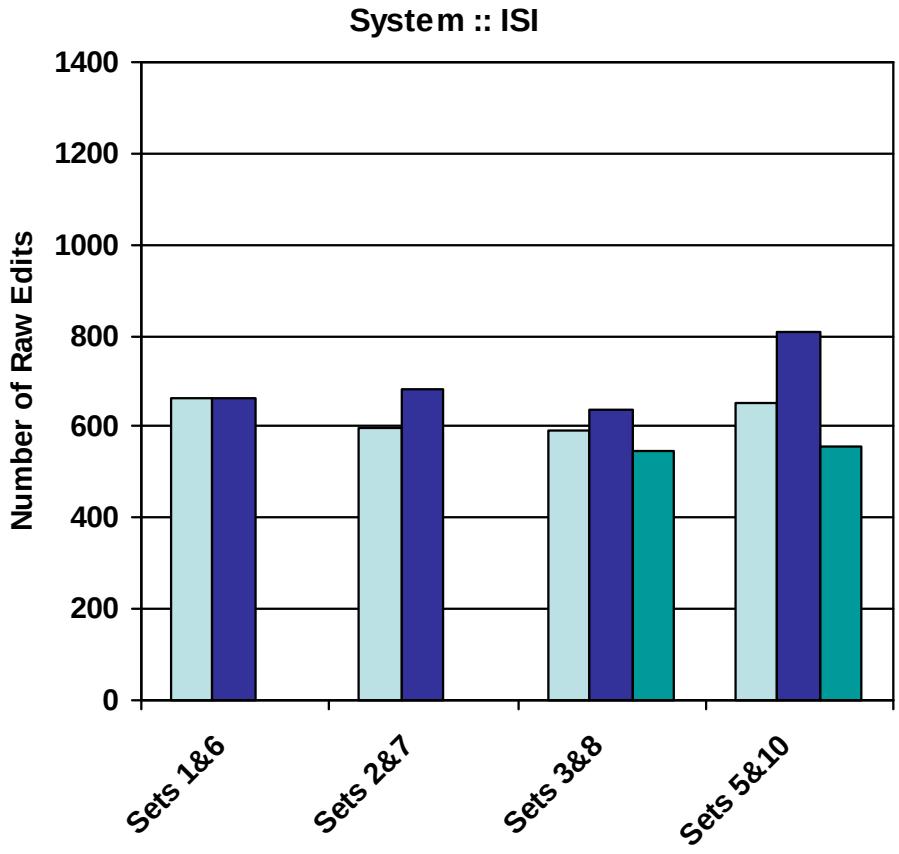
- Raw counts of edits over *30 common documents* (measured by TER) for each editor
  - Very similar “total edits” across editors for ISI system data



POC-2: Editor Agreement

- Raw counts of edits over *10 common documents* (measured by TER) of paired editors





- We see the same differences in number of total edits of between 150-200 edits

Metric Correlation with Human Judgment

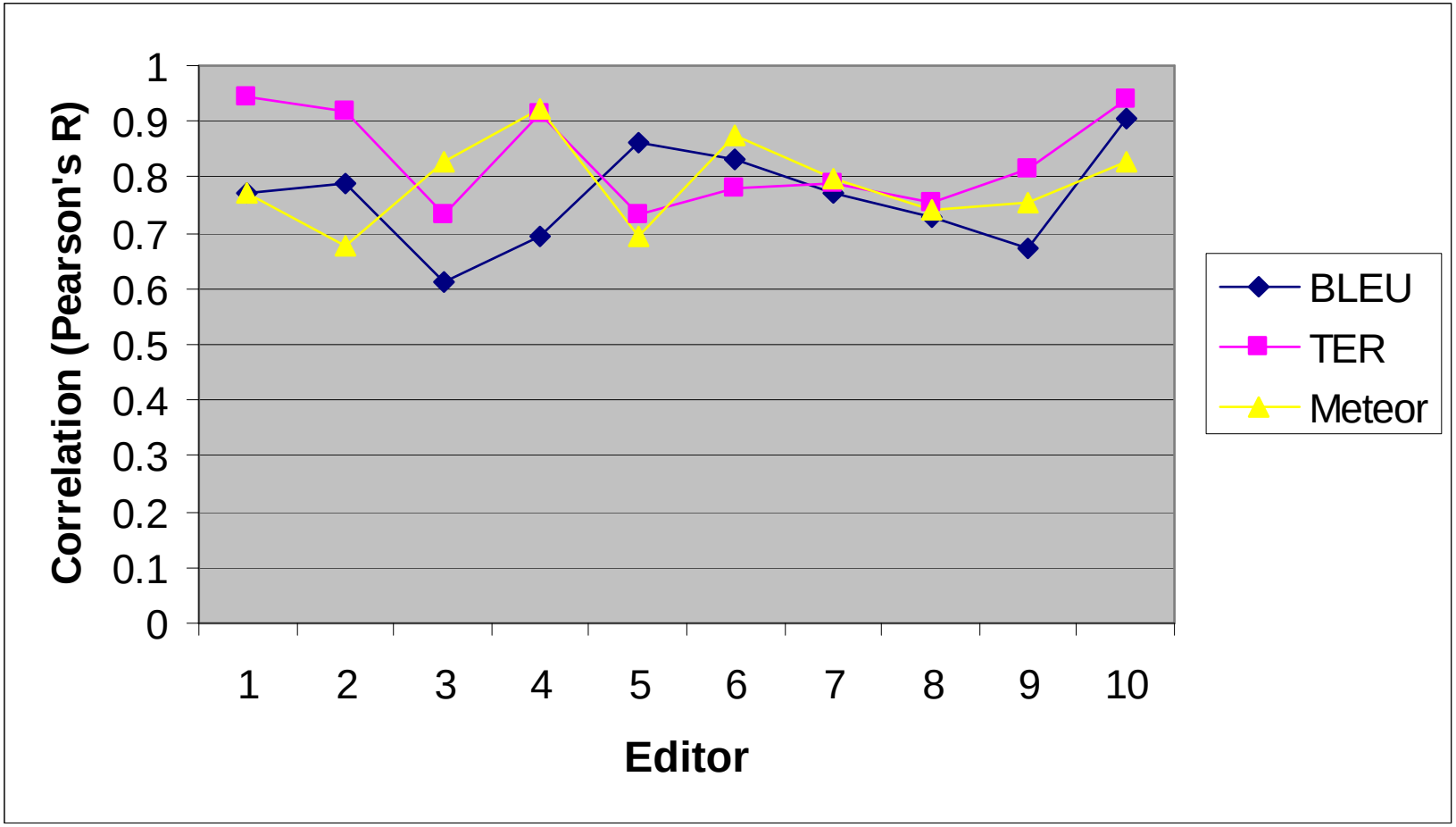
- Greg Sanders made a careful judgment of adequacy for each segment, which wasn't tainted by fluency
  - Created a document score by averaging segments
- Average correlation across editors between Greg's accuracy judgments and the different metrics

	Avg	StDev	StdErr
TER	0.831	0.087	0.027
BLEU	0.764	0.089	0.028
Meteor	0.789	0.078	0.025

Difference from TER Avg	Diff/StErr
-	
-0.067	-2.45
-0.042	-1.53

- TER is 2.45 standard deviations better than BLEU
- TER is 1.53 standard deviations better then Meteor

Metric With Strongest Correlation Differs Across Editors



More Human Judgments

Co-occurrence Counts For Greg's Judgments by Segment

- High Fluency with Low Accuracy (lower-left corner of table) did *not* occur
- High Accuracy with Low Fluency (upper-right corner) *did* occur

Greg Fluency	Greg Accuracy 1	2	3	4	5	
1	1	13	9	0	0	23
2	4	68	89	38	1	200
3	0	14	71	111	18	214
4	0	0	8	36	19	63
5	0	0	2	17	25	44
	5	95	179	202	63	544

Greg’s Fluency Judgments correlate less well than the Accuracy judgments

Greg's accuracy judgment correlated with each (doc score)

	avg	stdev	stderr	diff	diff/stderr
BLEU:	0.7640	0.0894	0.0283	-0.0673	-2.4497
TER:	0.8313	0.0868	0.0275		
METEOR:	0.7892	0.0777	0.0246	-0.0421	-1.5330

Greg's fluency judgment correlated with each (doc score)

	avg	stdev	stderr	diff	diff/stderr
BLEU:	0.6522	0.1319	0.0417	-0.0375	-0.7704
TER:	0.6897	0.1537	0.0486		
METEOR:	0.6918	0.1213	0.0384	0.0022	0.0449

Greg's (fluency+accuracy)/2 correlated with each (doc score)

	avg	stdev	stderr	diff	diff/stderr
BLEU:	0.7577	0.0811	0.0256	-0.0566	-2.9530
TER:	0.8143	0.0606	0.0192		
METEOR:	0.7908	0.0529	0.0167	-0.0235	-1.2270

Proof-of-Concept #3

- Goals:
  - Repeat the exercise with “audio” as the input source (*transcription* + *translation*)
  - Expose unique challenges speech data will present to the GALE Translation evaluation paradigm
    - How to handle disfluencies in speech
    - No longer have a predefined one-to-one segment mapping between the MT output and reference file for post editing
  - Use POC-3 to prepare for a Translation Dry-Run evaluation

POC-3: Data Set

- Documents (broadcasts)
  - Arabic and Chinese audio
  - 1 hour of broadcast conversations (talk shows, interviews, call-in programs, and roundtable discussions)
- Reference
  - LDC provided one high quality reference transcription file for each broadcast conversation (native language transcriptions)
  - ***Currently****:* LDC has contracted for two sets of translations per broadcast
- Two GALE teams produced MT output with site defined segment based time stamps
  - BBN/ISI for both Arabic and Chinese
  - IBM for Arabic

POC-3: Data Pre-Processing

- Transliteration filtering
  - DARPA will provide transliteration resources
- New challenge -- alignment
  - Systems won't always put the segments boundaries in the same place as the reference translation
- Proposed approach
  - NIST will align the segments that share the most overlap in time
- *Note, new challenge to post editors: they will be confronted with "meaning" that is split and merged among segments that are relevant to the reference*

POC-3: Post Editing

- Post Editors:
  - NIST Volunteers
    - 4 editors
      - 2 edit all Arabic data
      - 2 edit all Chinese data
    - 10-15 hours of post editing each editor
  - GALE Research Teams

- Schedule to finish
  - NIST expects translations by the first week of October
  - Post editing to finish by November 1st

Demonstrate the Post Editor Interface

Translation Dry-Run Evaluation

- Essential to guarantee a smooth and successful formal go/no-go evaluation next Summer
- Will
  - Be identical in scope to the formal evaluation
  - Be required for all GALE participants
  - Be completed well in advance of the formal evaluation
- Not to
  - Be viewed as establishing a baseline of performance

- NIST ***evaluation plan*** online
  - http://www.nist.gov/speech/tests/gale/2006dr/doc.htm

Translation Dry-Run Evaluation Task and Conditions

- One task:
  - Translation
- Two conditions:
  - Arabic to English
  - Chinese to English
- Two data sources:
  - Audio
    - *Broadcast news & talk shows*
    - Unstructured input, UEM files identify areas of waveform to be translated
  - Text
    - *Newswire* and *News groups*
    - Structured input, formatted similar to past NIST MT evaluations

Translation Dry-Run Evaluation Data Set

- Equal amounts of each data source for each language (~10,500 reference words)
  - ~30 text documents for each language
    - 15 newswire documents, 5,250 words of news group data
  - ~90 minutes of Arabic broadcasts
    - 45 minutes broadcast news and 45 minutes of broadcast conversations
  - ~60 minutes of Chinese broadcasts
    - 30 minutes of broadcast news and 30 minutes of broadcast conversations
- Data formats are defined in the evaluation plan

Translation Dry-Run Evaluation Data Pre-Processing

- NIST to create the Gold-Standard reference translation *to be worked out with NVTC*
  - LDC to provide three independent high quality translations (for each language)
  - Will identify disagreements, alternatives and ambiguities
    - Native speaker(s) to decide best choice, acceptable alternatives
- Transliterations normalized
- Alignments created between system output and reference

Translation Dry-Run Evaluation Metric

- Will use TER

- Editors will participate in a ***well defined training session***
  - Guidelines are posted on the NIST web space http://www.nist.gov/speech/tests/gale/poc/doc.htm
  - How many?
    - Several (at least 5) editors needed
  - Who?
    - LDC volunteered two post editors
    - Qualifications: Native English speakers, College students/graduates who have majored in English, possibly teachers, Technical writers …

GALE Translation 2006 Evaluation Schedule

Date	Milestone / Event
Nov-28-2005	• Data selection finalized • Source audio and segment delimitated source text delivered to NIST
Jan-11-2006	•Dry-run data sets delivered to sites
Jan-26-2005	• Reference translations delivered to NIST
Jan-31-2006	• MT translations due at NIST
Feb. 2006	• Post editing occurs
Mar-07-2006	• Resulting post edits and scores sent back to participants
Mar. 2006	• One-Day meeting
July 2006	• Gale Translation Evaluation

Remaining Issues

- Proposed alignment scheme
  - Changes may affect data formats (attributes)
- Qualifications of the Post Editors

- Translation of disfluencies for speech data
- Dry-Run Data
  - How will the data be selected
- Year-to-year test set comparison
  - Mothballed systems?
  - Progress test set?